L2VPN Working Group
Internet-Draft
Intended Status: Historical

Himanshu Shah
Ciena Corp

Eric Rosen
Francois Le Faucheur
Giles Heron
Cisco Systems
June 4, 2014

IP-Only LAN Service (IPLS)
draft-ietf-l2vpn-ipls-14.txt

Status of this Memo

This document is not an Internet Standards Track specificaion; it is published for the historical record.

This document defines a Historic Document for the Internet community. This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc6348.

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

Copyright Notice

Abstract

A Virtual Private LAN Service (VPLS) is used to interconnect systems across a wide-area or metropolitan-area network, making it appear that they are on a private LAN. The systems which are interconnected may themselves be LAN switches. If, however, they are IP hosts or IP routers, certain simplifications to the operation of the VPLS are possible. We call this simplified type of VPLS an "IP-only LAN Service" (IPLS). In an IPLS, as in a VPLS, LAN interfaces are run in promiscuous mode, and frames are forwarded based on their destination MAC addresses. However, the maintenance of the MAC forwarding tables is done via signaling, rather than via the MAC address learning procedures specified in [IEEE 802.1D]. This draft specifies the protocol extensions and procedures for support of the IPLS service.

The original intent was to provide an alternate solution to VPLS for those PE routers that were not capable of learning MAC address through data plane. This became non-issue with newer hardware. The concepts put forth by this draft are still valuable and are adopted in one form or other by newer work such as Ethernet VPN in L2VPN Working Group and possible data center applications. At this point, no further action is planned to update this document and is published simply as a historic record of the ideas.

Conventions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT",

"SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119.

Table of Contents

1.0 Contributing Authors

This document is the combined effort of the following individuals and many others who have carefully reviewed this document and provided the technical clarifications.

K. Arvind	Fortress
Vach Kompella/Mathew Bocci	Alcatel/Lucent
Shane Amante	Apple

2.0 Overview

As emphasized in [VPLS], Ethernet has become popular as an access technology in Metropolitan and Wide Area Networks. [VPLS] describes how geographically dispersed customer LANs can be interconnected over a service provider's network. The VPLS service is provided by Provider Edge (PE) devices that connect Customer Edge (CE) devices. The VPLS architecture provides this service by incorporating bridging functions such as MAC address learning in the PE devices.

Provider Edge platforms are designed primarily to be IP routers, rather than to be LAN switches. To add VPLS capability to a PE router, one has to add MAC address learning capabilities, along with aging and other mechanisms native to Ethernet switches. This may be fairly complex to add to the forwarding plane architecture of an IP router. As discussed in [L2VPN-FWK], in scenarios where the CE devices are NOT LAN switches, but rather are IP hosts or IP routers, it is possible to provide the VPLS service without requiring MAC address learning and aging on the PE. Instead, a PE router has to have the capability to match the destination MAC address in a packet received from a CE to an outbound pseudowire. The requirements for the IPLS service are described in [L2VPN-REQTS]. The purpose of this document is to specify a solution optimized for IPLS.

IPLS provides a VPLS-like service using PE routers that are not designed to perform general LAN bridging functions. One must be willing to accept the restriction that an IPLS be used for IP traffic only, and not used to interconnect CE devices that are

themselves LAN switches. This is an acceptable restriction in many environments, given that IP is the predominant type of traffic in today's networks.

The original intent was to provide an alternate solution to VPLS for those PE routers that were not capable of learning MAC address through data plane. This became non-issue with newer hardware. The concepts put forth by this draft are still valuable and are adopted in one form or other by newer work such as Ethernet VPN in L2VPN Working Group and possible data center applications. At this point, no further action is planned to update this document and is published simply as a historic record of the ideas.

In IPLS, a PE device implements multi-point LAN connectivity for IP traffic using the following key functions:

1. CE Address Discovery: Each Provider Edge (PE) device discovers MAC address of the locally attached Customer Edge (CE) IP devices, for each IPLS instance configured on the PE device. In some configurations, PE also learns the IP address of the CE device (when performing ARP proxy functions, described later in the document).

2. Pseudowire (PW) for Unicast Traffic: For each locally attached CE device in a given IPLS instance, a PE device sets up a pseudowire (PW-LSP) to each of the other PEs that supports the same IPLS instance.

   For instance, if PEx and PEy both support IPLS I, and PEy is locally attached to CEa and CEb, PEy will initiate the setup of two pseudowires between itself and PEx. One of these will be used to carry unicast traffic from any of PEx's CE devices to CEa. The other will be used to carry unicast traffic from any of PEx's CE devices to CEb.

   Note that these pseudowires carry traffic only in one direction. Further, while the pseudowire implicitly identifies the destination CE of the traffic, it does not identify the source CE; packets from different source CEs bound to the same destination CE are sent on a single pseudowire.

3. Pseudowires for Multicast Traffic: In addition, every PE supporting a given IPLS instance will set up a special 'multicast pseudowire' to every other PE in that IPLS instance. If, in the above example, one of PEx's CE devices sends a multicast packet, PEx would forward the multicast packet to PEy on the special 'multicast' pseudowire. PEy would then send a

copy of that packet to CEa and a copy to CEb.

The 'multicast' pseudowire carries Ethernet frames of multicast/broadcast IP, ARP and ICMP (Inverse) Neighbor Discovery (ND/IND) packets for IPv6. Thus when a PE sends a multicast packet across the network, it sends one copy to each remote PE (supporting the given IPLS instance). If a particular remote PE has more than one CE device in that IPLS instance, the remote PE must replicate the packet and send one copy to each of its local CEs.

As with the pseudowires that are used for unicast traffic, packets travel in only one direction on these pseudowires, and packets from different sources may be freely intermixed.

4. Signaling: The necessary pseudowires can be set up and maintained using the LDP-based signaling procedures described in [PWE3-CONTROL].

   A PE may assign the same label to each of the unicast pseudowires that lead to a given CE device, in effect creating a multipoint-to-point pseudowire.

   Similarly, a PE may assign the same label to each of the 'multicast' pseudowires for a given IPLS instance, in effect creating a multipoint-to-point pseudowire.
   When setting up a pseudowire to be used for unicast traffic, the PE must also signal the MAC address of the corresponding CE device. It should also, optionally, advertise IP address of the local CE device, especially when ARP proxy function is configured or simply for operational management purposes. Similarly, for IPv6 support, PE may optionally advertise the IPv6 addresses of the local CE device.

5. ARP Packet Forwarding: ARP packets [ARP] are forwarded from attachment circuit (AC) to 'multicast' pseudowires in the Ethernet frame format as described by [PWE3-ETH]. Following rules are observed when processing ARP packets,
   a. Both broadcast (request) and unicast (response) ARP packets are sent over the 'multicast' pseudowire.
   b. When an ARP packet is received from an AC, the packet is copied to control plane for learning MAC address of the CE. Optionally, IP address is also learned to record the association of IP and MAC address.
   c. All Ethernet packets, including ARP packets, received from 'multicast' pseudowire are forwarded out to all the ACs

associated with the IPLS instance. These packets are not copied to control plane.

6. ICMP IPv6 ND/IND related Packet Forwarding: (Inverse) Neighbor Discovery (ND/IND) IPv6 packets from an AC are replicated and a copy is sent to other ACs and to 'multicast' PWs associated with the IPLS instance in the native Ethernet format, unchanged. A copy is also submitted to Control Plane to learn the MAC address and optionally corresponding IPv6 addresses.

7. Multicast IP packet forwarding: An IP Ethernet frame received from an AC is replicated to other ACs and the 'multicast' pseudowires associated with the IPLS instance. An IP Ethernet frame received from a 'multicast' pseudowire is replicated to all the egress ACs associated with the IPLS instance.

8. Unicast IP packet forwarding: An IP packet received from the AC is forwarded based on the MAC DA lookup in the forwarding table. If a match is found, the packet is forwarded to the associated egress interface. If the egress interface is unicast pseudowire, the packet is sent without MAC header. If the egress interface is a local AC the Ethernet frame is forwarded as such. An IP packet received from the unicast pseudowire is forwarded to egress AC with MAC header prepended. The MAC DA is derived from the forwarding table while MAC SA is the MAC address of the PE.

Both VPLS [VPLS] and IPLS require the ingress PE to forward a frame based on its destination MAC address. However, two key differences between VPLS and IPLS can be noted from the above description:

. In VPLS, MAC entries are placed in the FIB of the ingress PE as a result of MAC address learning (which occurs in the data plane) while in IPLS MAC entries are placed in the FIB as a result of pseudowire signaling operations (control plane).
. In VPLS, the egress PE looks up a frame's destination MAC address to determine the egress AC; in IPLS, the egress AC is determined entirely by the ingress PW-label.

The following sections describe the details of the IPLS scheme.

2.1 Terminology

IPLS	IP-only LAN service (a type of Virtual Private LAN Service that is restricted to IP traffic only).

mp2p PW — Multipoint-to-Point Pseudowire. A pseudowire that carries traffic from remote PE devices to a PE device that signals the pseudowire. The signaling PE device advertises the same PW-label to all remote PE devices that participate in the IPLS service instance. In IPLS, for a given IPLS instance, an mp2p PW used for IP unicast traffic is established by a PE for each CE device locally attached to that PE. It is a unidirectional tree whose leaves consist of the remote PE peers (which connect at least one AC associated with the same IPLS instance) and whose root is the signaling PE. Traffic flows from the leaves towards the root.

Multicast PW — Multicast/broadcast Pseudowire. A special kind of mp2p PW that carries IP multicast/broadcast traffic, all ARP frames and ICMP (I)ND frames for IPv6. In the IPLS architecture, for each IPLS instance supported by a PE, that PE device establishes exactly one multicast PW. Multicast PW uses Ethernet encapsulation.

Unicast PW — Unicast Pseudowire carries IP unicast packets. A PE creates unicast PW for each locally attached CE. The unicast PW uses IP Layer2 transport encapsulation.

CE — Customer Edge device. In this document, a CE is any IP node (host or router) connected to the IPLS LAN service.

Replication Tree — The collection of all multicast PWs and ACs that are members of an IPLS service instance on a given PE. When a PE receives a multicast/broadcast packet from an AC, the PE device sends a copy of the packet to every multicast pseudowire and AC of the replication tree, excluding the AC on which the packet was received. When a PE receives a packet from a multicast PW, the PE device sends a copy of the packet to all the ACs of the replication tree and never to other PWs.

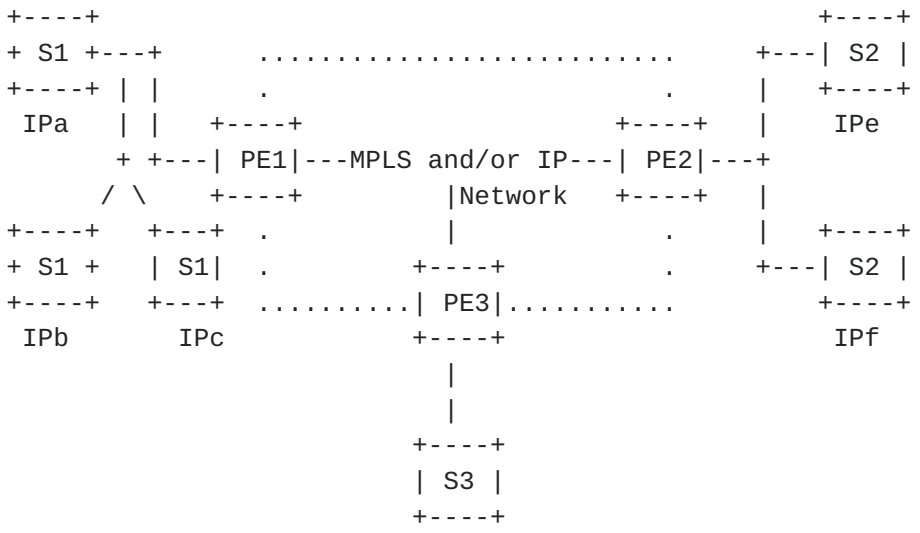
(I)ND — (Inverse) Neighbor Discovery in IPv6 uses ICMP

packets. It is a protocol that uses Neighbor solicitation/Advertisement PDUs.

RS	Router Solicitation. Hosts generate all router multicast ICMP packet to discover IPv6 router on the local link.
RA	Router Advertisement. Router generates all multicast ICMP packet to advertise its presence on the link. A unicast response is also sent when RS is received.
NS	Neighbor Solicitation in IPv6 uses (multicast) ICMP packets to resolve IPv6 interface address to MAC address association.
NA	Neighbor Advertisement in IPv6 uses (unicast) ICMP packets to respond to NS.

3.0 Topology

The Customer Edge (CE) devices are IP nodes (hosts or routers) that are connected to PE devices either directly, or via an Ethernet network. We assume that the PE/CE connection may be regarded by the PE as an "interface" to which one or more CEs are attached. This interface may be a physical LAN interface or a VLAN. The Provider Edge (PE) routers are MPLS Label Edge Routers (LERs) that serve as pseudowire endpoints.



In the above diagram, an IPLS instance is shown with three sites: site S1, site S2 and site S3. In site S3, the CE device is directly connected to its PE. In the other two sites, there are multiple CEs connected to a single PE. More precisely, the CEs at these sites are on an Ethernet (switched at site 1 and shared at site 2) network (or VLAN), and the PE is attached to that same Ethernet network or VLAN). We impose the following restriction: if one or more CEs attach to a PE by virtue of being on a common LAN or VLAN, there MUST NOT be more than one PE on that LAN or VLAN.

PE1, PE2 and PE3 are shown as connected via an MPLS network; however, other tunneling technologies, such as GRE, L2TPv3, etc., could also be used to carry the pseudowires.

An IPLS instance is a single broadcast domain, such that each IP end station (e.g., IPa) appears to be co-located with other IP end stations (e.g., IPb through IPf) on the same subnet. The IPLS service is transparent to the CE devices and requires no changes to them.

4.0 Configuration

Each PE router is configured with one or more IPLS service instances, and each IPLS service instance is associated with a unique VPN-Id. For a given IPLS service instance, a set of ACs is identified. Each AC can be associated with only one IPLS instance. An AC, in this document, is either a customer-facing Ethernet port, or a particular VLAN (identified by an IEEE 802.1Q VLAN ID) on a customer-facing Ethernet port.

The PE router can optionally be configured with a local MAC address to be used as source MAC address when IP packets are forwarded from

a pseudowire to an AC. By default, a PE uses the MAC address of the customer-facing Ethernet interface for this purpose.

5.0 Discovery

The discovery process includes:

- Remote PE discovery
- VPN (i.e., IPLS) membership discovery
- IP CE end station discovery

This draft does not discuss the remote PE discovery or VPN

membership discovery. This information can either be user configured or can be obtained using auto-discovery techniques described in [L2VPN-SIG] or other methods. However, the discovery of the CE is an important operational step in the IPLS model and is described below.

5.1 CE discovery

Each PE actively detects the presence of local CEs by snooping IP and ARP frames received over the ACs. When an AC configured in an IPLS instance becomes operational, it enters the CE discovery phase. In this phase, the PE examines each multicast/broadcast Ethernet frame. For link-local IP frames (for example IGP discovery/multicast/broadcast packets typically 224.0.0.x addresses [RFC-1112]), the CE's (source) MAC address is extracted from the Ethernet header and the (source) IP address is obtained from the IP header.

For each CE, the PE maintains the following tuple: <Attachment Circuit identification info, VPN-Id, MAC address, IP address (optional)>.

5.1.1 IPv4 based CE discovery

As indicated earlier, a copy of ARP frames received over the AC is submitted to the control plane. The PE learns MAC address and optionally IP address of the CE from the source address fields of the ARP PDU.

Once a CE is discovered, its status is monitored continuously by examining the received ARP frames and by periodically generating ARP requests. The absence of an ARP response from a CE after a configurable number of ARP requests is interpreted as loss of connectivity with the CE.

5.1.2 Ipv6 based CE discovery [RFC 4861]

A copy of Neighbor and Router Discovery frames received over the AC are submitted to the control plane in the PE.

If the PE receives a Neighbor Solicitation message, and the source IP address of the message is not the unspecified address, the PE learns the MAC address and optionally IP address of the CE.

If the PE receives an unsolicited Neighbor Advertisement message,

the PE learns the source MAC address and optionally the IP address of the CE.

If the PE receives a Router Solicitation, and the source IP address of the message is not the unspecified address, the PE learns source MAC address and optionally the IP address of the CE.

If the PE receives a Router Advertisement, it learns source MAC address and optionally the IP address of the CE.

The PE will periodically generate Neighbor Solicitation messages for the IP address of the CE as a means of verifying the continued existence of the address and its MAC address binding. The absence of a response from the CE device for a given number of retries could be interpreted as a loss of connectivity with the CE.

6.0 Pseudowire Creation

6.1 Receive Unicast Multipoint-to-point Pseudowire

As the PE discovers each locally attached CE, a unicast multipoint-to-point pseudowire (mp2p PW) associated exclusively with that CE is created by distributing the MAC address and optionally IP address of the CE along with a PW-Label to all the remote PE peers that participate in the same IPLS instance. Note that the same value of a PW-label SHOULD be distributed to all the remote PE peers for a given CE. The mp2p PW thus created is used by remote PEs to send unicast IP traffic to a specific CE.

(The same functionality can be provided by a set of point-to-point PWs, and the PE is not required to send the same PW-label to all the other PEs. For convenience, however, we will use the term mp2p PWs, which may be implemented using a set of point-to-point PWs.)

The PE forwards a frame received over this mp2p PW to the associated AC.

The unicast pseudowire uses IP Layer2 Transport encapsulation as define in [PWE3-CONTROL].

6.2 Receive Multicast Multipoint-to-point Pseudowire

When a PE is configured to participate in an IPLS instance, it advertises a 'multicast' PW-label to every other PE that is a member

of the same IPLS. The advertised PW-label value is the same for each

PE, which creates an mp2p pseudowire. There is only one such multicast mp2p PW per PE for each IPLS instance and this pseudowire is used exclusively to carry IP multicast/broadcast, ARP traffic and (inverse) Neighbor Discovery packets for IPv6 from the remote PEs to this PE for this IPLS instance.

Note that no special functionality is expected from this pseudowire. We call it a 'multicast' pseudowire because we use it to carry multicast and broadcast IP, ARP and IPv6 Neighbor Discovery traffic. The pseudowire itself need not provide any different service than any of the unicast pseudowires.

In particular, the Receive multicast mp2p PW does not perform any replication of frames itself. Rather, it is there to signify to the PE that the PE may need to replicate a copy of a frame received over this mp2p PW onto all the AC that are associated with the IPLS instance of the mp2p PW.

The multicast mp2p pseudowire is considered the principle pseudowire in the bundle of mp2p pseudowires that consist of one multicast mp2p pseudowire and a variable number of unicast mp2p pseudowires for a given IPLS instance. In a principle role, multicast PW represents the IPLS instance. The life of all unicast PWs in the IPLS instance depends on the existence of the multicast PW. If, for some reasons, multicast PW cease to exist, all the associated unicast pseudowires in the bundle are removed.

The multicast pseudowire uses Ethernet encapsulation as defined in [PWE3-ETH].

The use of pseudowires which are specially optimized for multicast is for further study.

6.3 Send Multicast Replication tree

The PE creates a send replication tree for each IPLS instance, which consists of the collection of all ACs and all the 'multicast' pseudowires of the IPLS instance.

Any ARP, Neighbor Discovery or multicast IP Ethernet frame received over an AC is replicated to the other ACs and to the mp2p multicast pseudowire of the send replication tree. The send replication tree deals mostly with broadcast/multicast Ethernet MAC frames. One exception to this is unicast ARP and IPv6 Neighbor Discovery frame, the processing of which is described in the following section.

Any Ethernet frame received over the multicast PW is replicated to all the ACs of the send replication tree of the IPLS instance associated with the incoming PW label. One exception is unicast ARP

and Neighbor Discovery frame used for IPv6, the processing of which is described in the following section.

7.0 Signaling

[PWE3-CONTROL] uses the Label Distribution Protocol (LDP) to exchange PW-FECs in the Label Mapping message in a downstream unsolicited mode. The PW-FEC comes in two forms; PWid and Generalized PWid FEC elements. These FEC elements define some fields that are common between them. The discussions below refer to these common fields for IPLS related extensions. Note that the use of multipoint to point and unidirectional characteristics of the PW makes BGP as the ideal candidate for PW-FEC signaling. The use of BGP for such purposes is for future study.

7.1 IPLS PW Signaling

An IPLS carries IP packets as payload over its unicast pseudowires and Ethernet packet as payload over its multicast pseudowire. The PW-type to be used for unicast pseudowire is the IP PW, defined in [PWE3-CONTROL] as IP Layer2 Transport. The PW-type to be used for multicast pseudowire is the Ethernet PW as defined in [PWE3-ETH]. The PW-Type values for these encapsulations are defined in [PWE3-IANA].

When processing a received PW FEC, the PE matches the PW Id with the locally configured PW Id for the IPLS instance. If the PW type is Ethernet, the PW-FEC is for multicast PW. If the PW type is 'IP Layer2 transport', the PW FEC is for unicast PW.

For unicast PW, PE must check the presence of MAC address TLV in the optional parameter fields of the Label Mapping message. If this parameter is absent, a Label Release message must be issued with a Status Code meaning "MAC Address of the CE is absent" [note: Status Code 0x000000XX is pending IANA allocation], to reject the establishment of the unicast PW with the remote PE.

The PE may optionally include IP address TLV based on the user configuration for advertising of the IP addresses of the local CE.

The processing of the address list TLV is as follows.

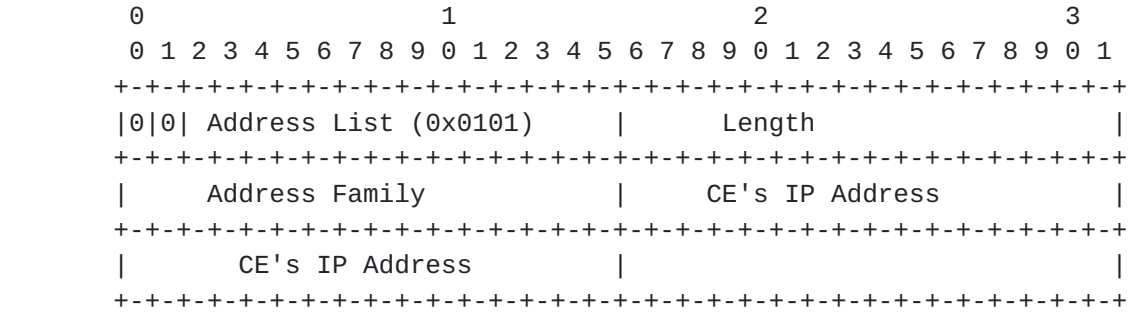
- If a pseudowire is configured for AC with IPv4 CEs only, the PE should advertise address list tlv with address family type to be of IPv4 address. The PE should process the IPv4 address list TLV as described in this document.
- If a pseudowire is configured for AC with both IPv4 and IPv6 CEs, the PE should advertise IPv6 capability using the procedures described in Section below.

  o If a PE does not receive any IP address list TLV or IPv6
     capability advertisement, it MAY assume IPv4 behavior.

The IPLS uses the Address List TLV as defined in [RFC 5036] to signal the MAC (and optionally IP) address of the local CE. There are two TLVs defined below; IP Address TLV and MAC Address TLV. MAC address TLV must be included in the optional parameter field of the Label Mapping message when establishing the unicast IP PW for IPLS.

When configured to support specific type of IP traffic (IPv4 or IPv6), the PE augments verification of the type of traffic PW will carry using the Address Family Type value. If there is a mismatch between the received Address Family value and the expectation of IPLS instance to which the PW belongs, the PE must issue a Label Release message with a Status Code meaning "IP Address type mismatch" (Status Code 0x0000004A) to reject the PW establishment.

Encoding of the IP Address TLV is:



Length
  When Address Family is IPV4, Length is equal to 6 bytes;
  2 bytes for address family and 4 bytes of IP address.

Address Family
  Two octet quantity containing a value from the ADDRESS FAMILY NUMBERS from ADDRESS FAMILY NUMBERS in [RFC 3232] that encodes the addresses contained in the Addresses field.

IP Address of the CE
  IP address of the CE attached to the advertising PE. The encoding of the individual address depends on the Address Family.

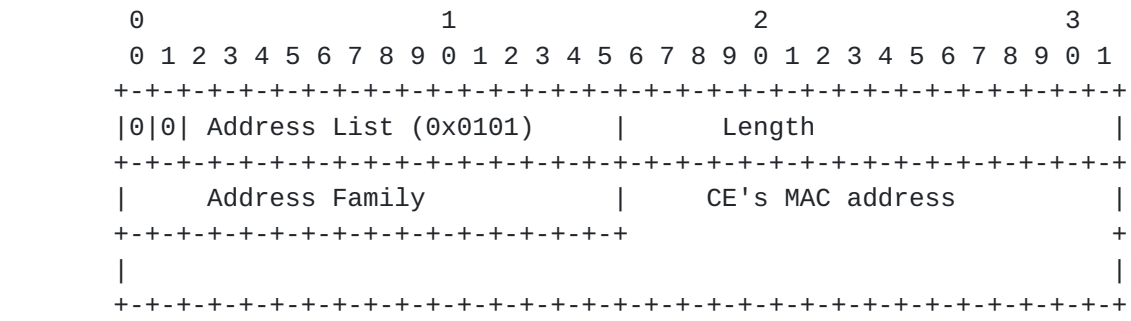
The following address encodings are defined by this version of the protocol:

Address Family	Address Encoding
IPv4 (1)	4 octet full IPv4 address

IPv6 (2)	16 octet full IPv6 address

Note that more than one instance of the IP address TLV may exist, especially when support for IPv6 is configured.

Encoding of the MAC Address TLV is:



Length
The length field is set to value 8 (2 for address family, 6 for MAC address)

Address Family
Two octet quantity containing a value from ADDRESS FAMILY NUMBERS in [RFC 3232] that encodes the addresses contained in the Addresses field.

CE's MAC Address
MAC address of the CE attached to the advertising PE. The encoding of the individual address depends on the Address Family.

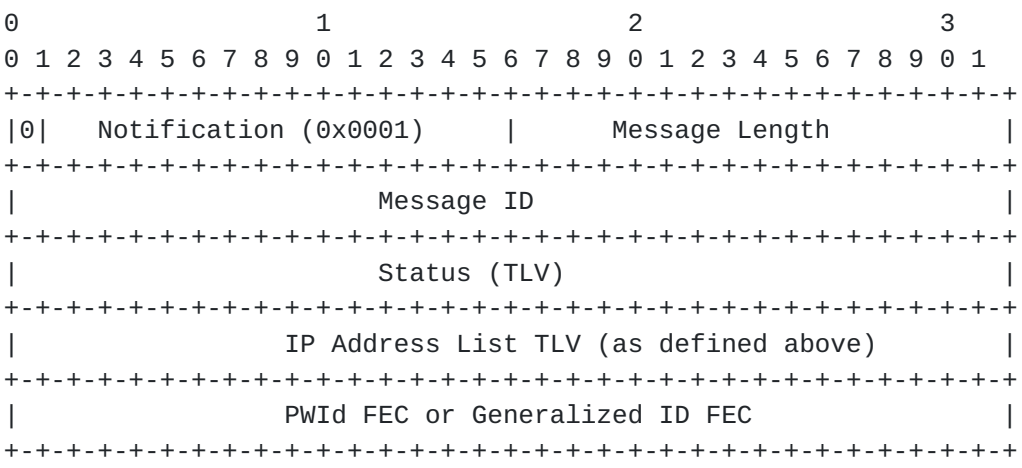
The following address encodings are defined by this version of the protocol:

Address Family	Address Encoding
MAC (6)	6 octet full Ethernet MAC address

The IPv4 address of the CE is also supplied in the optional parameters field of the LDP Notification message along with the PW FEC. The LDP Notification message is used to signal any change in the status of the CE's IPv4 address.

Note that Notification message does not apply to MAC address TLV since an update to MAC address of the CE should result in label withdraw followed by establishment of new PW with new MAC address of the CE. However, advertisement of IP address(es) of the CE is optional and changes may become known after the establishment of unicast PW.

The encoding of the LDP Notification message is as follows.



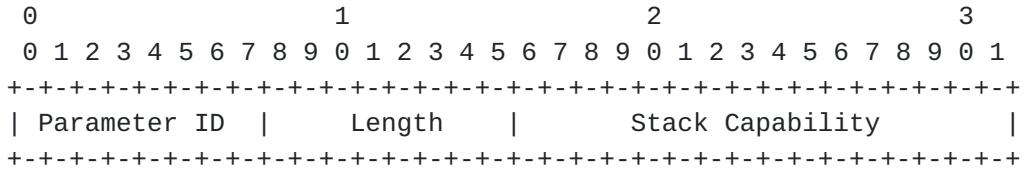
The Status TLV status code is set to 0x0000002C "IP address of CE", to indicate that IP Address update follows. Since this notification does not refer to any particular message the Message Id, and Message Type fields are set to 0.

The PW FEC TLV SHOULD NOT include the interface parameters as they are ignored in the context of this message.

7.2 IPv6 Capability Advertisement

A 'Stack Capability' Interface Parameter sub-TLV is signaled by the two PEs so that they can agree which stack(s) they should be using. It is assumed by default that the IP PW will always be capable of carrying IPv4 packets. Thus this capability sub-TLV is used to indicate if other stacks need to be supported concurrently with IPv4.

The 'Stack Capability' sub-TLV is part of the interface parameters of the PW FEC. The proposed format for the Stack Capability interface parameter sub-TLV is as follows:



Parameter ID = 0x16

Length = 4

Stack capability = 0x000X to indicate IPv6 stack capability

The Value of Stack capability is dependent on the PW type context. For IP PW type, a setting of 0x000X indicates IPv6 stack capability.

A PE that supports IPv6 on an IP PW MUST signal the stack capability sub-TLV in the initial label mapping message for the PW. The PE nodes compare the value advertised by the remote PE with the local configuration and only use a capability which is advertised by both. If a PE that supports IPv6 does not receive a 'stack capability' sub-TLV from the far-end PE in the initial label mapping message, or one is received but it is set to a reserved value, the PE MUST send an unsolicited release for the PW label with the LDP status code meaning "IP Address type mismatch" (Status Code 0x0000004A).

The behavior of a PE that does not understand an interface parameter sub-TLV is specified in RFC4447 [PWE3-CONTROL].

7.3 Signaling Advertisement Processing

A PE should process a received [PWE3-CONTROL] advertisement with PW-type of IP Layer2 transport for IPLS as follows,

- Verify the IPLS VPN membership by matching the VPN-Id signaled in the AGI field or the PW-ID field with all the VPN-Ids configured in the PE. Discard and release the PW label if VPN-Id is not found.
- Program the Forwarding Information Base (FIB) such that when a unicast IP packet is received from an AC with its destination MAC address matching the advertised MAC address, the packet is forwarded out over the tunnel to the

advertising PE with the advertised PW-label as the inner label.

A PE should process a received [PWE3-CONTROL] advertisement with the PW type of Ethernet for IPLS as follows,

- Verify the IPLS VPN membership by matching the VPN-Id signaled in the AGI field or the PW-ID field with all the VPN-Ids configured in the PE. Discard and release the PW label if VPN-Id is not found.
- Add the PW-label to the send broadcast replication tree for the VPN-Id. This enables sending a copy of a multicast/broadcast IP Ethernet frame or ARP Ethernet frame or Neighbor Discovery frames from the AC to this pseudowire.

8. IANA Considerations

Since this document is being published as historic record, no requests for IANA code points are necessary. However, if in future, interest to pursue this proposal arises, the following requests for IANA codes would become necessary.

8.1. LDP Status messages

This document uses new LDP status code. IANA already maintains a registry of name "STATUS CODE NAME SPACE" defined by [RFC 5036]. The following value is suggested for assignment:

0x000000XX "MAC Address of CE is absent"

8.2. Interface Parameters

This document proposes a new Interface Parameters sub-TLV, to be assigned from the 'Pseudowire Interface Parameters Sub-TLV type Registry'. The following value is suggested for the Parameter ID:

0xXX "Stack capability"

IANA is also requested to set up a registry of "L2VPN PE stack capabilities". This is a 16 bit field. Stack capability values 0x000X is specified in Section 7. of this document. The remaining bitfield values (0x0002,..,0x8000) are to be assigned by IANA using the "IETF Consensus" policy defined in [RFC 5226].

L2VPN PE Stack Capabilities:

```
Bit (Value)       Description
===============   ========================================
Bit 0  (0x000X) - IPv6 stack capability
Bit 1  (0x000X) - Reserved
Bit 2  (0x000X) - Reserved
          .
          .
          .
Bit 14 (0xX000) - Reserved
Bit 15 (0xX000) - Reserved
```

9.0 Forwarding

9.1 Non-IP or non-ARP traffic

In an IPLS VPN, a PE forwards only IP and ARP traffic. All other frames are dropped silently. If the CEs must pass non-IP traffic to each other, they must do so through IP tunnels that terminate at the CEs themselves.

9.2 Unicast IP Traffic

In IPLS, IP traffic is forwarded from the AC to the PW based on the destination MAC address of the layer 2 frame (and not based on the IP Header).

The PE identifies the FIB associated with an IPLS instance based on the AC or the PW label. When a frame is received from an AC, the PE uses the destination MAC address as the lookup key. When a frame is received from a PW, the PE uses the PW-Label as the lookup key. The frame is dropped if the lookup fails.

For IPv6 support, the unicast IP ICMP frame of Neighbor Discovery Protocol [RFC 4861] is bi-casted; one copy is submitted to the control plane and other copy to the PW, based on the destination MAC address.

9.3 Broadcasts and Multicast IP Traffic

When the destination MAC address is either a broadcast or multicast, a copy of the frame is sent to the control plane for CE discovery purposes (see section 5.1). It is important to note that the frames sent to the control plane is applied stricter rate limiting criteria to avoid overwhelming the control plane under adverse conditions

such as Denial Of Service attack. The service provider should also provide a configurable limitation to prevent overflowing of the learned source addresses in a given IPLS instance. Also, a caution must be used such that only link local multicasts and broadcast IP packets are sent to control plane.

When a multicast/broadcast IP packet is received from an AC, the PE replicates it onto the Send Multicast Replication Tree (See section 6.3). When a multicast/broadcast IP Ethernet frame is received from a pseudowire, the PE forwards a copy of the frame to all the ACs associated with the respective IPLS VPN instance. Note that 'multicast' PW uses Ethernet encapsulation and hence does not require additional header manipulations.

9.4 ARP Traffic

When a broadcast ARP frame is received over the AC, a copy of the frame is sent to the control plane for CE discovery purposes. The PE replicates the frame onto the Send Multicast Replication Tree (see section 6.3), which results into a copy to be delivered to all the remote PEs on the 'multicast' PW and other local CEs through the egress ACs.

When a broadcast Ethernet ARP frame is received over the 'multicast' PW, a copy of the Ethernet ARP frame is sent to all the ACs associated with the IPLS instance.

When a unicast Ethernet ARP frame is received over the AC, a copy of the frame is sent to the control plane for the CE discovery purposes. The PE may optionally do MAC DA lookup in the forwarding table and send the ARP frame to a specific egress interface (AC or 'multicast' PW to a remote PE) or replicate the frame onto the Send Multicast Replication Tree (see section 6.3).

When a unicast ARP Ethernet frame is received over the 'multicast' PW, PE may optionally do MAC DA lookup in the forwarding table and forward it to the AC where the CE is located. If the CE is not accessible through any local AC, the frame is dropped. Conversely, the PE may simply forward the frame to all the ACs associated with that IPLS instance without any lookup in the forwarding table.

9.5 Discovery of IPv6 CE devices

A PE device that supports IPv6 MUST be capable of,

- Intercepting ICMPv6 Neighbor Discovery [RFC 4861] packets

received over the AC.
- Record the IPv6 interface addresses and CE link-layer addresses present in these packets
- Forward them towards the original destination

A PE device may also intercept Router Discovery packets in order to discover the link layer address and IPv6 interface address(es) of the CE. Following sections describe the details.

The PE device MUST learn the link-layer address of the local CE and be able to use it when forwarding traffic between CEs. The PE MAY also wish to monitor the source link-layer address of data packets received from the CE, and discard packets not matching its learned CE link-layer address. The PE device may also optionally learn a list of CE IPv6 interface addresses for its directly-attached CE.

9.5.1. Processing of Neighbor Solicitations

When a broadcast Neighbor Solicitation frame is received over the AC, a copy of the frame is sent to the control plane for CE discovery purposes. The PE replicates the frame onto the Send Multicast Replication Tree (see section 6.3), which results into a copy to be delivered to all the remote PEs on the 'multicast' PW and other local CEs through the egress ACs. The PE may optionally learn an IPv6 interface address (If provided - this will not be the case for Duplicate Address Detection) when present.

When a broadcast Ethernet Neighbor Solicitation frame is received over the 'multicast' PW, a copy is sent to all the ACs associated with the IPLS instance.

9.5.2 Processing of Neighbor Advertisements

When a unicast Neighbor Advertisement is received over the AC, a copy of the frame is sent to the control plane for the CE discovery purposes. The PE may optionally do MAC DA lookup in the forwarding table and send the Neighbor Advertisement frame to a specific egress interface (AC or 'multicast' PW to a remote PE) or replicate the frame onto the Send Multicast Replication Tree (see section 6.3).

Optionally, PE could learn the IPv6 Interface address of the CE.

When a unicast Neighbor Advertisement frame is received over the 'multicast' PW, PE may optionally do MAC DA lookup in the forwarding table and forward it to the AC where the CE is located. If the CE is not accessible through any local AC, the frame is dropped.

Conversely, the PE may simply forward the frame to all the ACs associated with that IPLS instance without any lookup in the forwarding table.

9.5.3 Processing of Inverse Neighbor Solicitations and Advertisement

Inverse Neighbor Discovery is typically used on non-broadcast links, but are allowed on broadcast links too [RFC 3122]. PE may optionally intercept Inverse Neighbor Solicitation and Advertisement and learn MAC and IPv6 interface address list of the attached CE from the copy of the frame sent to the control plane. The PE may optionally do MAC DA lookup in the forwarding table and send another copy of the frame to a specific egress interface (AC or 'multicast' PW to a remote PE) or replicate the frame onto the Send Multicast Replication Tree (see section 6.3).

9.5.4 Processing of Router Solicitations and Advertisements

Router Solicitations (RS) are multicast while Router Advertisement (RA) can be unicast or multicast Ethernet frames. The PE could optionally intercept RS and RA frames and send a copy to control plane. The PE may learn the MAC address and a list of interface addresses for the attached CE.

For unicast RA, the PE may optionally do MAC DA lookup in the forwarding table and send the Neighbor Advertisement frame to a specific egress interface (AC or 'multicast' PW to a remote PE) or replicate the frame onto the Send Multicast Replication Tree (see section 6.3). The multicast RA and RS Ethernet frames are replicated to using the Send Multicast Replication Tree as described in section 6.3.

9.6 Encapsulation

The Ethernet MAC header of a unicast IP packet received from an AC is stripped before forwarding the frame to the unicast pseudowire. However, the MAC header is retained for the following cases,

- when a frame is unicast or broadcast IP packet that is directed to one or more local AC(s).
- when a frame is a broadcast IP packet
- when a frame is an ARP packet
- when a frame is Neighbor/Router Solicitation/Advertisement

An IP frame received over a unicast pseudowire is prepended with a MAC header before transmitting it on the appropriate ACs). The

fields in the MAC header are filled in as follows:

- The destination MAC address is the MAC address associated with the PW label in the FIB
- The source MAC address is the PE's own local MAC address or a MAC address which has been specially configured on the PE for this use.
- The Ethernet Type field is 0x0800 if IPv4 or 0x86DD if IPv6 [RFC 2464]
- The frame may be IEEE802.1Q tagged based on the VLAN information associated with the AC.

An FCS is appended to the frame.

10.0 Attaching to IPLS via ATM or FR

In addition to (i) an Ethernet port and a (ii) combination of Ethernet port and a VLAN ID, an AC to IPLS may also be (iii) an ATM or FR VC carrying encapsulated bridged Ethernet frames or (iv) the combination of an ATM or FR VC and a VLAN ID.

The ATM/FR VC is just used as a way to transport Ethernet frames between a customer site and the PE. The PE terminates the ATM/FR VC and operates on the encapsulated Ethernet frames exactly as if those were received on a local Ethernet interface. When a frame is propagated from pseudowire to a ATM or FR VC the PE prepends the Ethernet frame with the appropriate bridged encapsulation header as defined in [RFC 2684] and [RFC 2427] respectively. Operation of an IPLS over ATM/FR VC is exactly as described above, with the exception that the AC is then identified via the ATM VCI/VPI or Frame Relay DLCI (instead of via a local Ethernet port ID), or a combination of those with a VLAN ID.

11.0 VPLS vs IPLS

The VPLS approach proposed in [VPLS] provides VPN services for IP as well as other protocols. The IPLS approach described in this draft is similar to VPLS in many respects:

- It provides a Provider Provisioned Virtual LAN service with multipoint capability where a CE connected via a single attachment circuit can reach many remote CEs
- It appears as a broadcast domain and a single subnet
- forwarding is based on destination MAC addresses

However, unlike VPLS, IPLS is restricted to IP traffic only. By restricting the scope of the service to the predominant type of traffic in today's environment, IPLS eliminates the need for service

provider edge routers to implement some bridging functions such as MAC address learning in the data path (by, instead, distributing MAC information in the control plane). Thus this solution offers a number of benefits:

- Facilitates Virtual LAN services in instances where PE devices cannot or cannot efficiently (or are specifically configured not to) perform MAC address learning.
- Unknown Unicast frames are never flooded as would be the case in VPLS.
- Encapsulation is more efficient (MAC header is stripped) for unicast IP packets while traversing the backbone network.
- PE devices are not burdened with the processing overhead associated with traditional bridging (e.g., STP processing, etc.). Note however that some of these overheads (e.g., STP processing) could optionally be turned-off with a VPLS solution in the case where it is known that only IP devices are interconnected.
- Loops (perhaps through backdoor links) are minimized since a PE could easily reject (via label release) a duplicate IP to MAC address advertisement.
- Greater control over CE topology distribution.

12.0 IP Protocols

The solution described in this document offers IPLS service for IPv4 and IPv6 traffic only. For this reason, the MAC Header is not carried over the unicast pseudowire. It is reconstructed by the PE when receiving a packet from a unicast pseudowire and the Ethertype 0x0800 or 0x86DD is used in the MAC Header since IPv4 or IPv6 respectively, is assumed.

However, this solution may be extended to carry other types of important traffic such as ISIS , which does not use Ethernet-II, EtherType based header. In order to permit the propagation of such packets correctly, one may create a separate set of pseudowires, or pass protocol information in the "control word" of a "multiprotocol" pseudowire, or encapsulate the Ethernet MAC Header in the pseudowire. The selection of appropriate multiplexing/demultiplexing scheme is the subject of future study. The current document focuses on IPLS service for IPv4 and IPv6 traffic.

13.0 Dual Homing with IPLS

As stated in previous sections, IPLS prohibits connection of a common LAN or VLAN to more than one PE. However, the CE device

itself can connect to more than one instance of IPLS through two separate LAN or VLAN connections to separate PEs. To the CE IP device, these separate connections appear as connections to two IP subnets. The failure of reachability through one subnet is then resolved via the other subnet using IP routing protocols.

14.0 Proxy ARP function

The earlier version of this proposal used IP-PW to carry both the broadcast/multicast and unicast IP traffic. It also discussed how PE proxy functionality responds to the ARP requests of the local CE on behalf of remote CE. The current version of the draft eliminated these functions and instead uses Ethernet PW to carry broadcast, multicast and ARP frames to remote PEs. The motivation to use Ethernet PW and propagate ARP frames in the current version is to support configuration like back-to-back IPLS (similar to Inter AS option-A configurations in [RFC 4364]).

The termination and controlled propagation of ARP frames is still a desirable option for security, DoS and other purposes. For these reasons, we re-introduce the ARP Proxy [PROXY-ARP] function in this revision as an optional feature. Following sections describe this option.

14.1 ARP Proxy - Responder

As a local configuration, a PE can enable ARP Proxy responder function. In this mode, local PE responds to ARP requests received over the Attachment Circuit via learnt IP and MAC address associations, which are advertised by the remote PEs. In addition, PE may utilize local policies to determine if ARP requests should be responded based on the source of the ARP request, rate at which the ARP requests are generated, etc. In nutshell, when this feature is enabled, ARP requests are not propagated to remote PE routers that are members of the same IPLS instance.

14.2 ARP Proxy - Generator

As a local configuration, a PE can enable ARP Proxy generator function. In this mode, the PE generates ARP request for each IP and MAC address associations received from the remote PEs. The remote CE's IP and MAC address is used as the source information in the ARP request while the destination IP address in the request is obtained from the local configuration (that is, user needs to configure an IP address when this feature is enabled). The ARP request is sent on

the Attachment Circuits that have ARP Proxy Generator enabled and is

associated with the given IPLS instance.

In addition, the PE may utilize local policies to determine which IP/MAC addresses are candidate for ARP request generation.

The ARP Proxy Generator feature is required to support back-to-back IPLS configuration when any member of the IPLS instance is using ARP Proxy Responder function. An example of a back-to-back IPLS is a configuration where PE-1 (ASBR) in an IPLS cloud in one Autonomous System (say, AS-1) is connected via an Attachment Circuit to another PE-2 (ASBR) in an IPLS cloud in another Autonomous System (say, AS-2) where each PE appears as CE to each other. Such configuration is described in [RFC 4364] as option-A for inter-AS connectivity. The Proxy ARP responder feature prevents propagation of ARP requests to PE-1 (ASBR) in AS-1. This necessitates that PE-1 (ASBR) in AS-1 generate ARP request on behalf of each CE connected to the IPLS instance in AS-1 as a mean to 'advertise' the reachability to IPLS cloud in AS-2

15.0 Data Center Applicability

The resurgence of interest in providing IP/MPLS based solution for Data Center Networks (DCN) deserves another look at the IPLS methodologies described in this document. The key requirement of DCN to permit VM mobility within or across DCN necessiates extending the reachability of IP subnet over a LAN, transparently. In addition, VMs tendancy to generate frequent gratutious ARPs for location discovery necessiates a solution that curbs broadcasts closest to the source.

The IPLS solution facilitates VM mobility by way of PE closest to the new location signaling the MAC address to all remote peers. In addition, control-plane based MAC learning mechanisms prevent flooding of unknown unicast across DCN. The optional ARP proxy mechanisms further reduces ARP broadcast floods by preventing its reach across local PE.

16.0 Acknowledgements

Authors would like to thank Alp Dibirdi from Alcatel, Xiahou from Huawei and other L2VPN working group members for their valuable comments.

17.0 Security Considerations

A more comprehensive description of the security issues involved in L2VPNs are covered in [VPN-SEC]. Most of the security issues can be

avoided through implementation of appropriate guards. The security aspect of this solution is addressed for two planes; control plane and data plane.

17.1 Control plane security

The control plane security pertains to establishing the LDP connection, pseudo-wire establishment and CE's IP and MAC address distribution. The LDP connection between two trusted PEs can be achieved by each PE verifying the incoming connection against the configured peer's address and authenticating the LDP messages using MD5 authentication. The pseudo-wire establishments between two secure LDP peers do not pose security issue but mis-wiring could occur due to configuration error. Some checks, such as, proper pseudo-wire type and other pseudo-wire options may prevent mis-wiring due to configuration errors.

The learning of the appropriate CE's IP and MAC address can be a security issue. It is expected that the local attachment circuit to CE be physically secured. If this is a concern, the PE must be configured with CE's IP and MAC address. During each ARP frame processing, PE must verify the received information against the configuration before accepting. This prevents theft of service, denial of service to a subscriber or DoS attacks to all subscribers by malicious use of network services.

The IPLS also provides MAC anti spoofing by preventing the use of already known MAC address. For instance, if a PE has already learned a presence of a CE through local connection or from another PE, and subsequently an advertisement for the same MAC and/or IP address is received from a different PE, the receiving PE can terminate service to that CE (either through label release and/or removing the ARP entry from the FIB) and raise the alarm.

The IPLS learns and distributes CE reachability through the control plane. This provides greater control over CE topology distribution through application of local policies.

17.2 Data plane security

The data traffic between CE and PE is not encrypted and it is possible that in an insecure environment, a malicious user may tap into the CE to PE connection and generate traffic using the spoofed destination MAC address on the Ethernet Attachment Circuit. In order to avoid such hijacking, local PE may verify the source MAC address of the received frame against the MAC address of the

admitted connection. The frame is forwarded to PW only when authenticity is verified. When spoofing is detected, PE must severe the connection with the local CE, tear down the PW and start over.

Each IPLS instance uses its own FIB. This prevents leaking of one customer data into another.

18.0 References

18.1 Normative References

[ARP] RFC 826, STD 37, D. Plummer, "An Ethernet Address Resolution Protocol".

[PWE3-CONTROL] L. Martini et al., "Pseudowire Setup and Maintenance using LDP", RFC 4447.

[PWE3-IANA] L. Martini et al,. "IANA Allocations for pseudo Wire Edge to Edge Emulation (PWE3)", RFC 4446.

[PWE3-ETH] Martini et al., "Encapsulation Methods for Transport of Ethernet over MPLS Networks", RFC 4448.

[VPLS] Lasserre et al, "Virtual Private LAN Service Using LDP", RFC 4762, January 2007.

[RFC 5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "LDP Specification", RFC 5036, October 2007.

[IEEE 802.1D] ISO/IEC 10038, ANSI/IEEE Std 802.1D-1993, "MAC Bridges".

[RFC 4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.

[RFC 2464] Crawford, M., "Transmission of IPv6 packets over Ethernet Networks", RFC 2464, December 1998.

[RFC 3122] Conta, A., "Extensions to IPv6 Neighbor Discovery for Inverse Discovery Specification", RFC 3122, June 2001.

[RFC 5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226,

May 2008.

18.2 Informative References

[L2VPN-FWK] Andersson, L., Ed., and E. Rosen, Ed., "Framework for Layer 2 Virtual Private Networks (L2VPNs)", RFC 4664, September 2006.

[PROXY-ARP] RFC 925, J. Postel, "Multi-LAN Address Resolution".

[L2VPN-REQTS] Augustyn, W. et.al "Service Requirements for Layer 2 Provider Provisioned Virtual Private Networks", RFC 4665, September 2006.

[L2VPN-SIG] Rosen et al., "Provisioning, Autodiscovery, and signaling in L2VPN", RFC 6074, Jan 2011.

[RFC-1112] Deering, S., "Host Extensions for IP Multicasting", RFC 1112, August, 1989.

[RFC 2684] Grossman, et al., "Multiprotocol Encapsulation over ATM Adaptation Layer 5", September 1999.

[RFC 2427] Brown, et al., "Multiprotocol Interconnect over Frame Relay", September 1998.

[RFC 4364] Rosen et al., "BGP/MPLS IP Virtual Private Network (VPNs)", February 2006.

[VPN-SEC] Fang, L., "Security framework for Provider Provisioned Virtual Private Networks", RFC 4111, July 2005.

[RFC 3232] Reynolds and Postel, "Assigned Numbers".

18.0 Author's Address

Himanshu Shah

Ciena Corp
3939 North 1st Street,
San Jose, CA 95110
Email: hshah@ciena.com

Eric Rosen
Cisco Systems
300 Apollo Drive,
Chelmsford, MA 01824
Email: erosen@cisco.com

Giles Heron
Cisco Systems
Email: giheron@cisco.com

Francois Le Faucheur
Cisco Systems, Inc.
Village d'Entreprise Green Side - Batiment T3
400, Avenue de Roumanille
06410 Biot-Sophia Antipolis, France
Email: flefauch@cisco.com